

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2001-273214

(43)Date of publication of application : 05.10.2001

(51)Int.Cl. G06F 13/00
G06F 17/21

(21)Application number : 2000-083150

(71)Applicant : OKI SOFTWARE KK
OKI ELECTRIC IND CO LTD
NTT ME CORP

(22)Date of filing : 24.03.2000

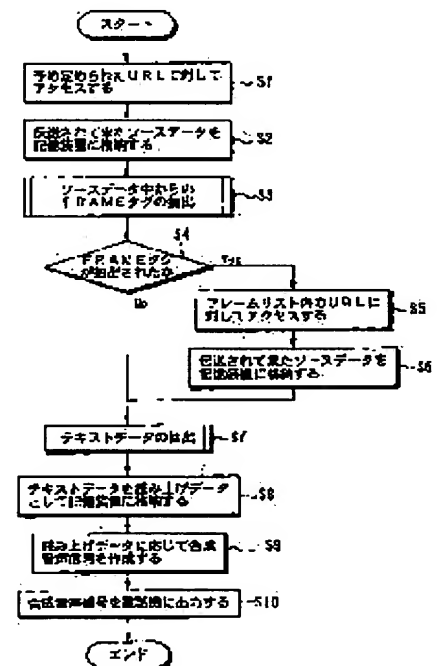
(72)Inventor : NISHIMURA HITOSHI
YAKIDA KAZUHIKO
MATSUSHITA ARIYUKI
YAMAGUCHI YUICHIRO
ITO SHINICHI
NAGAI TOMOYASU
OTA TSUYOSHI
TAMURA MASARU

(54) WEB PAGE DECODING SYSTEM

(57)Abstract:

PROBLEM TO BE SOLVED: To provide a Web page decoding system capable of exactly extracting a text part in respect to the document of an HTML containing a tag having the description of a uniform resource locator(URL).

SOLUTION: Basic source data comprising a Web page are extracted from a storage area designated by the prescribed URL, and written in a storage means and when the existence of the tag containing the description of the URL is detected out of the basic source data, the URL in that prescribed tag is detected. Then, source data are extracted from the storage area designated by that detected URL, and written in the storage means and the text part is extracted from all the source data stored in the storage means.



(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開2001-273214

(P2001-273214A)

(43) 公開日 平成13年10月5日 (2001. 10. 5)

(51) Int.Cl. ⁷	識別記号	F I	テーマコード*(参考)
G 0 6 F 13/00	3 5 4	G 0 6 F 13/00	3 5 4 D 5 B 0 0 9
17/21	5 0 1	17/21	5 0 1 T 5 B 0 8 9
	5 6 8		5 6 8 A
	5 9 6		5 9 6 A

審査請求 未請求 請求項の数6 O L (全 6 頁)

(21) 出願番号 特願2000-83150 (P2000-83150)

(22) 出願日 平成12年3月24日 (2000. 3. 24)

(71) 出願人 591051645

沖ソフトウェア株式会社

東京都板橋区舟渡1丁目12番11号

(71) 出願人 000000295

沖電気工業株式会社

東京都港区虎ノ門1丁目7番12号

(71) 出願人 596094692

株式会社エヌ・ティ・ティ エムイー

東京都千代田区大手町二丁目2番2号

(74) 代理人 100079119

弁理士 藤村 元彦

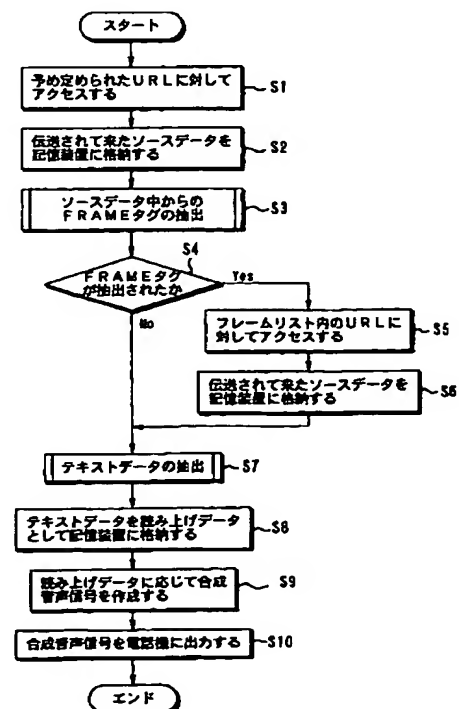
最終頁に続く

(54) 【発明の名称】 ウェブページ解読システム

(57) 【要約】

【課題】 URL (ユニホームリソースロケータ) の記述を有するタグを含む HTML の文書に対してテキスト部分を正確に抽出することができる Web (ウェブ) ページ解読システムを提供する。

【解決手段】 Web ページを構成する基本のソースデータを所定の URL で指定された記憶領域から取り出して記憶手段に書き込み、基本のソースデータ中から URL の記述箇所を含む所定のタグの存在を検出した場合には、その所定のタグ中の URL を検出し、その検出した URL で指定された記憶領域からソースデータを取り出して記憶手段に書き込み、記憶手段に記憶されたソースデータの全てからテキスト部分を抽出する。



【特許請求の範囲】

【請求項1】 Web（ウェブ）ページを構成するHTML文書のテキスト部分を解読するWebページ解読システムであって、

前記Webページを構成する基本のソースデータを所定のURL（ユニホームリソースロケータ）で指定された記憶領域から取り出して記憶手段に書き込む手段と、

前記基本のソースデータ中からURLの記述箇所を含む所定のタグの存在を検出するタグ検出手段と、

前記所定のタグの存在が検出された場合にはその所定のタグ中のURLを検出するURL検出手段と、

前記URL検出手段によって検出されたURLで指定された記憶領域からソースデータを取り出して前記記憶手段に書き込む手段と、

前記記憶手段に記憶されたソースデータの全てからテキスト部分を抽出するテキスト抽出手段と、を備えたことを特徴とするウェブページ解読システム。

【請求項2】 前記所定のタグはフレームタグであることを特徴とする請求項1記載のウェブページ解読システム。

【請求項3】 前記URL検出手段は、前記フレームタグ中のURLとして<FRAME SRC="URL">の構文中のURLを検出することを特徴とする請求項1又は2記載のウェブページ解読システム。

【請求項4】 前記テキスト抽出手段は、前記記憶手段に記憶されたソースデータ中の<>で囲まれた部分以外の部分をテキスト部分として抽出することを特徴とする請求項1記載のウェブページ解読システム。

【請求項5】 前記テキスト抽出手段によって抽出されたテキスト部分に対応して音声信号を作成して出力する音声出力手段を更に備えたことを特徴とする請求項1記載のウェブページ解読システム。

【請求項6】 前記音声出力手段の出力音声信号は公衆電話回線を介して電話機に供給されることを特徴とする請求項1記載のウェブページ解読システム。

【発明の詳細な説明】

【0001】

【発明が属する技術分野】 本発明は、Web（ウェブ）ページのテキスト部分を解読するWebページ解読システムに関する。

【0002】

【従来の技術】 インターネットの情報サービスの1つであるWWW(World Wide Web)は、HTML(Hyper Text Markup Language)という言葉で記述されたHTMLファイルとそのファイルの保存位置の識別子であるURL(Uniform Resource Locator)とを用いてインターネットを介して文字、映像、音声等のマルチメディア情報を参照することができるものである。HTMLファイルをWWWブラウザと呼ばれる閲覧ソフトウェアによって処理することによりディスプレイ画面上に形成されるものがW

ebページである。WWWの情報を提供する側、すなわちWWWサーバはHTMLファイルをURLで関連付けて保存しており、サーバアプリケーションに従って動作する。情報を提供される側、すなわちクライアント（クライアントコンピュータ）ではWWWブラウザを用いて所望のURLからHTMLファイルを含むソースデータ（例えば、画像ファイルや音声ファイル）にインターネットを介してアクセスしてソース中のファイルによって形成されたWebページをディスプレイ画面上で参照することができる。

【0003】 HTMLの文書は、通常、テキスト文書からなるテキスト部とタグによって形成される表示情報部とから形成される。タグは<>を一對とする記号であり、タグを用いてHTMLの構文が例えば、<HTML>〜</HTML>の如く形成される。タグ<>で囲まれた部分にはWebページに表示されるテキスト部の文字の大きさ、フォントの種類、その文字色、Webページの背景色、画像ファイル名、画像位置等の様々な表示情報が示される。

20 【0004】 このようにHTMLの文書においてはタグ<>で囲まれた部分は表示される部分ではなく、文書を表示するための制御情報であるので、タグ<>で囲まれた部分を除くと、通常、単なるテキスト文書となることが普通である。一方、WWWブラウザに表示されるWebページの文書を読み上げるシステムがインターネット上には形成されることがある。これは、クライアントの端末がコンピュータではなく、例えば、公衆回線に接続された電話機である場合にWebページの文書を読み上げて音声信号として電話機に送出するためである。Webページ読上としては読み上げ対象のHTMLファイル中からタグ<>で囲まれた部分を除くテキストデータ部分を抽出し、そのテキストデータ部分の文字コードに対応した音声データを合成して一連の音声信号として出力することが行われる。

【0005】

【発明が解決しようとする課題】 HTMLには、Webページ上にフレームを形成するための構文として、例えば、<FRAMESET>〜</FRAMESET>があり、これを用いたWebページでは分割された画面が得られる。その構文が記述されたHTMLファイルからは分割画面毎に別のHTMLファイルが更に呼び出されて文書が表示される。すなわち、<FRAME SRC="URL">の如きタグにより分割画面毎にURL（HTMLファイル名を含む）が指定され、その指定されたURLの領域に存在するHTMLファイルの内容が表示される。

40 【0006】 しかしながら、従来、このようなフレームタグのようにURLの記述を有するタグを含むHTMLの文書に対してはテキスト部分を正確に抽出することができないという問題点があった。そこで、本発明の目的

は、URLの記述を有するタグを含むHTMLの文書に対してテキスト部分を正確に抽出することができるWebページ解読システムを提供することである。

【0007】

【課題を解決するための手段】本発明のWebページ解読システムは、Webページを構成するHTML文書のテキスト部分を解読するWebページ解読システムであって、Webページを構成する基本のソースデータを所定のURL(ユニホームリソースロケータ)で指定された記憶領域から取り出して記憶手段に書き込む手段と、基本のソースデータ中からURLの記述箇所を含む所定のタグの存在を検出するタグ検出手段と、所定のタグの存在が検出された場合にはその所定のタグ中のURLを検出するURL検出手段と、URL検出手段によって検出されたURLで指定された記憶領域からソースデータを取り出して記憶手段に書き込む手段と、記憶手段に記憶されたソースデータの全てからテキスト部分を抽出するテキスト抽出手段と、を備えたことを特徴としている。この構成より、基本のソースデータがフレームタグを含むHTMLファイルの場合には、そのフレームタグ内に記述されたURLの領域に格納されているHTMLファイルのテキスト部分も抽出することができる。

【0008】

【発明の実施の形態】以下、本発明の実施例を図面を参照しつつ詳細に説明する。図1は本発明のによるWebページ解読システムの構成を示している。このシステムにおいては、WWWサーバ1は情報サービスとしてWWWを提供するサーバであり、HTMLファイルをURLに関連付けて保存しており、また、画像ファイルや音声ファイルも保存している。WWWサーバ1はインターネット回線網2に接続されている。

【0009】インターネット回線網2にはCTI(Computer Telephony Integration)サーバ3が接続されている。CTIサーバ3は公衆電話回線網4にも接続されている。公衆電話回線網4には複数の電話機が実際には接続されているが、ここでは1つの電話機5を示している。電話機は一般加入電話機、公衆電話機及び携帯電話機のいずれであっても良い。なお、公衆電話回線網4には中継局、基地局等の電話回線接続のための局が存在するが、図には示していない。

【0010】CTIサーバ3はWebページの読み上げを電話機5を含む電話機のユーザに提供するサーバである。CTIサーバ3には、Webページ取得部31と、テキスト抽出部32と、テキスト読み上げ部33とが備えられている。Webページ取得部31はWWWサーバ1にアクセスし、Webページのソースデータを取得する。テキスト抽出部32はWebページ取得部31によって取得されたソースデータを解析し、テキスト部分を抽出する。テキスト読み上げ部33はテキストの文字コードに応じて合成音声信号を作成し、その合成音声信号

を公衆電話回線網4を利用して電話機に対して出力する。Webページ取得部31、テキスト抽出部32及びテキスト読み上げ部33はCTIサーバ3のプロセッサ(図示せず)の後述の如き動作によって形成される。

【0011】また、CTIサーバ3は内部にハードディスク等の記憶装置35を有しており、後述するように、ソースデータ等の各種データが記憶装置35には記憶される。WWWサーバ1及びCTIサーバ3各々のインターネット回線網2を利用した通信においては通信プロトコルとしてTCP/IPが用いられ、WWWサーバ1及びCTIサーバ3にはIPアドレスが各々割り当てられている。更に、WWWのプロトコルとしてはHTTPが使用される。また、図示していないが、WWWサーバ1及びCTIサーバ3はルータを介してインターネット回線網2には接続されている。

【0012】次に、かかるWebページ解読システムの動作について説明する。ユーザが電話機5からCTIサーバ3へ電話をかけ、電話機5とCTIサーバ3との間の通話状態が確立すると、CTIサーバ3は図2に示すように、まず、予め定められたURLで指定される領域のWebページのソースデータを取得するために、そのURLに対してアクセスを行う(ステップS1)。このURLがWWWサーバ1内にあるとすると、WWWサーバ1はURLで指定される領域のHTMLファイル等のファイルからなるソースデータを読み出してCTIサーバ3に対して送信する。そのソースデータはWebページを構成する基本となるソースデータである。送信されたソースデータはインターネット回線網2を介してCTIサーバ3に供給される。

【0013】CTIサーバ3はWWWサーバ1から送られて来たソースデータを記憶装置35に格納し(ステップS2)、その格納したソースデータ中からFRAME(フレーム)タグを抽出する(ステップS3)。このFRAMEタグの抽出動作については後述するが、FRAMEタグ中に含まれるURLがフレームリストとして記憶装置35に書き込まれる。

【0014】CTIサーバ3はステップS3の実行の結果として、FRAMEタグの抽出が行われたか否かを判別する(ステップS4)。ステップS4にてFRAMEタグの抽出が実際に行われなかった場合には、後述のステップS7に進む。一方、FRAMEタグの抽出が実際に行われた場合には、フレームリストのURLで指定される領域のWebページのソースデータを取得するために、そのフレームリストのURLに対してアクセスを行い(ステップS5)、WWWサーバ1から送られて来たソースデータを記憶装置35に格納する(ステップS6)。ステップS5のアクセスに対するWWWサーバ1の動作はステップS1のアクセスの場合と同様である。ステップS6の実行後はステップS7に進む。

【0015】CTIサーバ3はステップS7において記

憶装置 35 に格納されたソースデータからタブ<>で囲まれた部分以外のテキスト部分を抽出し、その抽出テキストデータを読み上げデータとして記憶装置 35 に書き込む(ステップ S 8)。その後、読み上げデータに基づいて合成音声信号を作成し(ステップ S 9)、その合成音声信号を電話機 5 に対して出力する(ステップ S 10)。記憶装置 35 に書き込まれた読み上げデータは複数の文字コードからなるテキストデータであるので、その文字コード各々又は単語単位の文字コード群に対応する音声データを記憶装置 35 から検索して得て、それら音声データを合成して連続する合成音声信号を作成する。合成音声信号は公衆電話回線網 4 を介して電話機 5 に供給され、電話機 5 の受話器から読み上げ音出力される。なお、記憶装置 35 には文字コードと音声データとの関係を示すデータテーブルが予め記憶されている。

【0016】次に、上記のステップ S 3 におけるソースデータ中からの FRAME タグ抽出動作について図 3 のフローチャートを参照しつつ説明する。CTI サーバ 3 は、記憶装置 35 に記憶されたソースデータ中から文字列<FRAME SRCを検索し(ステップ S 11)、文字列<FRAME SRCがソースデータ中に存在するか否かを判別する(ステップ S 12)。すなわち、記憶装置 35 に書き込まれたソースデータ中には HTML ファイルが含まれ、その HTML ファイルが示す文書でフレーム設定が行われているか否かが判別される。文字列<FRAME SRCが存在するならば、<FRAME SRC="URL">の構文が存在するので、読み取り位置をその次の文字" "の位置まで移動し(ステップ S 13)、更に、その後の" "で囲まれた文字列、すなわち URL をソースデータから読み取り、その URL を記憶装置 35 に形成されたフレームリストに書き込む(ステップ S 14)。よって、フレームリストにはフレーム内に含まれる HTML 文書の存在位置を示す URL が書き込まれる。ステップ S 14 の実行後、ソースデータの全てのファイルから文字列<FRAME SRCの検索が終了したか否かを判別し(ステップ S 15)、その検索が終了していない場合にはステップ S 11 に戻り、上記のステップ動作を繰り返す。

【0017】次いで、上記したステップ S 7 におけるテキストデータの抽出動作について図 4 のフローチャートを参照しつつ説明する。CTI サーバ 3 は、記憶装置 35 に格納されたソースデータのうちの 1 つのファイルの先頭から順に 1 文字分の文字コードを取得し(ステップ S 21)、その文字コードが文字<を示すか否かを判別する(ステップ S 22)。取得した文字コードが文字<を示す場合にはタグフラグ F_{TAG} を 1 に等しくさせる(ステップ S 23)。

取得した文字コードが文字<を示さない場合にはタグフラグ F_{TAG} が 1 に等しいか否かを判別する(ステップ S 24)。タグフラグ F_{TAG} は HTML ファイルにおいて<>で囲まれた部分において 1 に

設定され、それ以外の部分において 0 に設定されるフラグであり、その初期値は 0 である。ステップ S 23 の実行後もステップ S 24 の判別は実行される。

【0018】ステップ S 24 の判別の結果、タグフラグ F_{TAG} が 1 に等しくされている場合には、ステップ S 21 で取得した文字コードが文字>を示すか否かを判別する(ステップ S 25)。取得した文字コードが文字>を示す場合にはタグの終了であるので、タグフラグ F_{TAG} を 0 に等しくさせる(ステップ S 26)。一方、ステップ S 24 の判別の結果、タグフラグ F_{TAG} が 0 に等しくされている場合には、<>で囲まれたタグ部分以外のテキスト部分であるので、取得した文字コードを読み上げデータに含ませるように記憶装置 35 に格納する(ステップ S 27)。

【0019】ステップ S 26 又は S 27 の実行後はソースデータの全てから 1 文字分の文字コードの取得が終了したか否かを判別し(ステップ S 28)、その取得が終了していない場合にはステップ S 21 に戻り、上記のステップ動作を繰り返す。よって、かかるシステムによれば、基本のソースデータが FRAME タグを含む HTML ファイルの場合に FRAME タグ内に記述された URL の領域に格納されている HTML ファイルのテキスト部分も抽出することができるので、WWW ブラウザに表示される Web ページのテキスト部分を余すことなく読み上げることができる。

【0020】上記した実施例においては、フレームタグが使用された場合について説明したが、フレームタグ以外の URL の記述を有するタグにも本発明を適用することができる。また、JavaScript 等のスクリプト言語を含む HTML ファイルからテキスト部分を抽出して読み上げる場合にも本発明を適用することができる。HTML では、機能を拡張するために Web ページ上で JavaScript 等のスクリプト言語を実行できるようにするタグも用意されている。例えば、<SCRIPT LANGUAGE="JavaScript">~</SCRIPT>のような構文で形成される。よって、上記したように<SCRIPT LANGUAGE="JavaScript">~</SCRIPT>の範囲の部分を無視してそれ以外のテキスト部分を抽出するのである。

【0021】更に、本発明のシステムは HTML を用いたファイルの場合に限らず、HTML を拡張させた言語を用いたファイルについても適用することができる。なお、日本では WWW ブラウザにて閲覧できるページを全てホームページと称しているが、ホームページは本来、1 情報群を構成する複数の Web ページのうちの基本ページであるので、ここでは誤解を招かないように Web ページと記載した。

【0022】

【発明の効果】以上の如く、本発明の Web ページ解説

10

20

30

40

50

システムにおいては、URLの記述を有するタグを含むHTMLの文書に対してテキスト部分を正確に抽出することができる。よって、Webページを読み上げる際にはWebページのテキスト部分を余すことなく読み上げることができる。

【図面の簡単な説明】

【図1】本発明によるWebページ解読システムの構成を示すブロック図である。

【図2】図1のシステム中のCTIサーバの動作を示すフローチャートである。

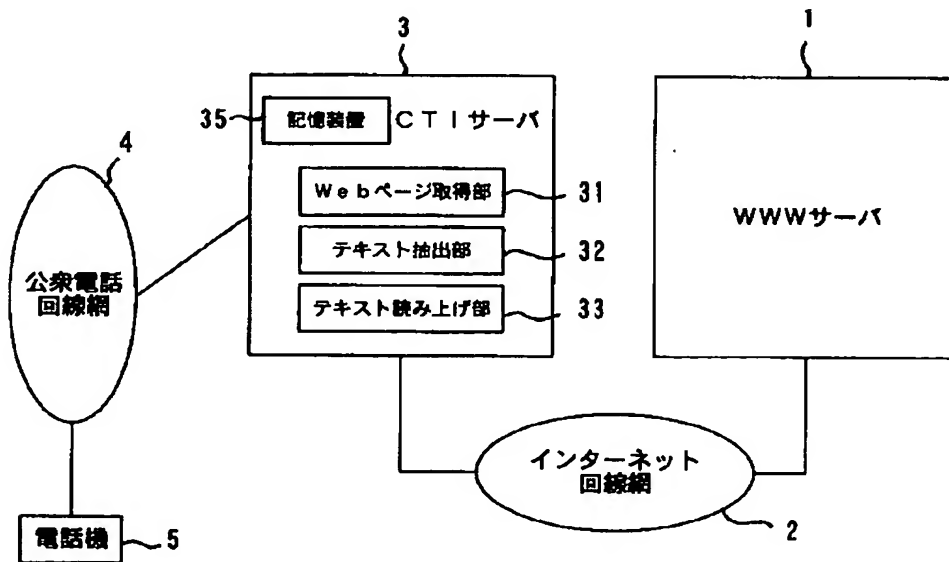
*【図3】ソースデータ中からのFRAMEタグ抽出動作を示すフローチャートである。

【図4】テキストデータの抽出動作を示すフローチャートである。

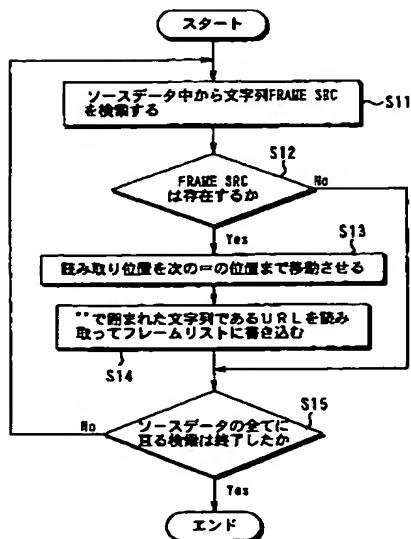
【符号の説明】

- 1 WWWサーバ
- 2 インターネット回線網
- 3 CTIサーバ
- 4 公衆電話回線網
- 5 電話機

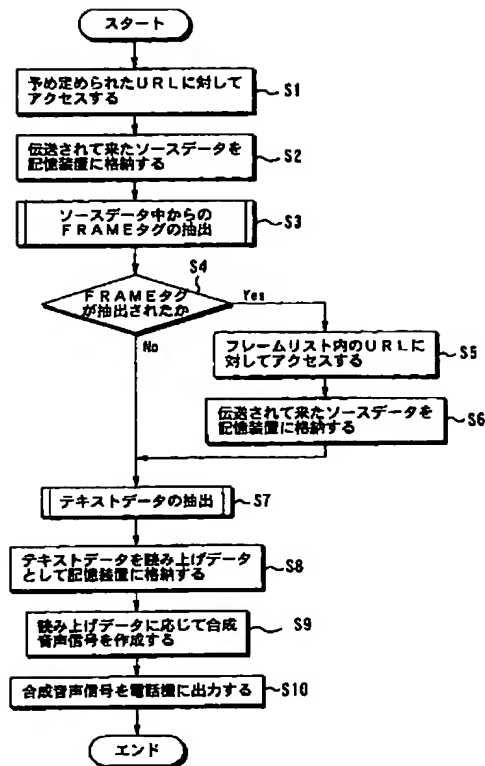
【図1】



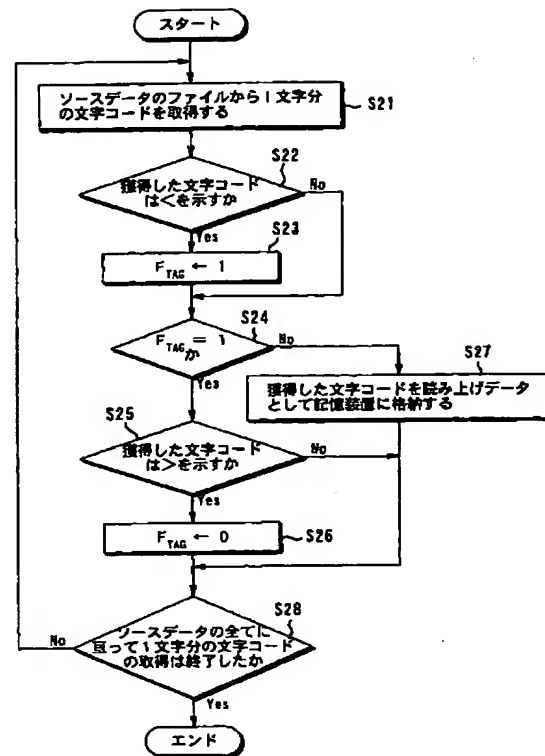
【図3】



【図2】



【図4】



フロントページの続き

- (72)発明者 西村 仁司
東京都板橋区舟渡1丁目12番11号 沖ソフトウェア株式会社内
- (72)発明者 八木田 一彦
東京都板橋区舟渡1丁目12番11号 沖ソフトウェア株式会社内
- (72)発明者 松下 有亨
東京都板橋区舟渡1丁目12番11号 沖ソフトウェア株式会社内
- (72)発明者 山口 雄一郎
東京都港区虎ノ門1丁目7番12号 沖電気工業株式会社内
- (72)発明者 伊藤 慎一
東京都港区虎ノ門1丁目7番12号 沖電気工業株式会社内

- (72)発明者 永井 友康
東京都千代田区大手町2-2-2 アーバンネット大手町ビル 株式会社エヌ・ティ・ティエムイー内
- (72)発明者 大田 剛志
東京都千代田区大手町2-2-2 アーバンネット大手町ビル 株式会社エヌ・ティ・ティエムイー内
- (72)発明者 田村 賢
東京都千代田区大手町2-2-2 アーバンネット大手町ビル 株式会社エヌ・ティ・ティエムイー内
- Fターム(参考) 5B009 QA11 RD03 SA03 SA14 TA11
VA02 VC01
5B089 GA11 GB03 HA01 JA22 JB02
KA04 KB07 KC53 KC59 LB13

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2001-273214

(43)Date of publication of application : 05.10.2001

(51)Int.Cl.

G06F 13/00
G06F 17/21

(21)Application number : 2000-083150

(71)Applicant : OKI SOFTWARE KK
OKI ELECTRIC IND CO LTD
NTT ME CORP

(22)Date of filing : 24.03.2000

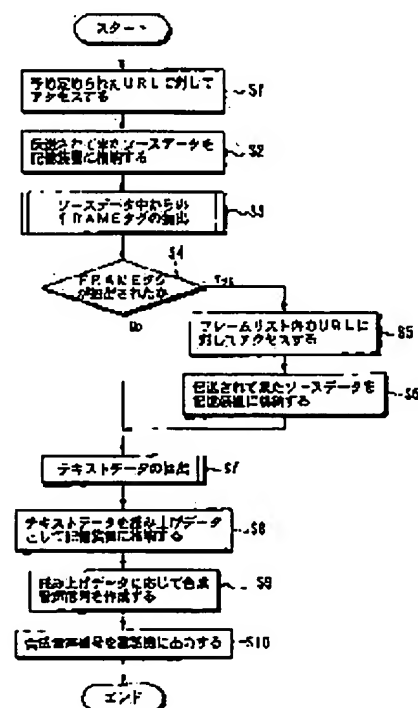
(72)Inventor : NISHIMURA HITOSHI
YAKIDA KAZUHIKO
MATSUSHITA ARIYUKI
YAMAGUCHI YUICHIRO
ITO SHINICHI
NAGAI TOMOYASU
OTA TSUYOSHI
TAMURA MASARU

(54) WEB PAGE DECODING SYSTEM

(57)Abstract:

PROBLEM TO BE SOLVED: To provide a Web page decoding system capable of exactly extracting a text part in respect to the document of an HTML containing a tag having the description of a uniform resource locator (URL).

SOLUTION: Basic source data comprising a Web page are extracted from a storage area designated by the prescribed URL, and written in a storage means and when the existence of the tag containing the description of the URL is detected out of the basic source data, the URL in that prescribed tag is detected. Then, source data are extracted from the storage area designated by that detected URL, and written in the storage means and the text part is extracted from all the source data stored



in the storage means.

* NOTICES *

JP0 and INPIT are not responsible for any damages caused by the use of this translation.

- 1.This document has been translated by computer. So the translation may not reflect the original precisely.
- 2.**** shows the word which can not be translated.
- 3.In the drawings, any words are not translated.

CLAIMS

[Claim(s)]

[Claim 1]A Web page decipherment system which decodes a text part of an HTML document which constitutes a Web (web) page, comprising:

A means which takes out source data of foundations which constitute said Web page from a storage area specified by predetermined URL (uniform resource locator), and is written in a memory measure.

A tag detection means which detects existence of a predetermined tag including a description part of URL out of source data of said foundations.

A URL detection means to detect URL in the predetermined tag when existence of said predetermined tag is detected.

A means which takes out source data from a storage area specified by URL detected by said URL detection means, and is written in said memory measure, and a sampling-of-text means to extract a text part from all the source data memorized by said memory measure.

[Claim 2]The web page decipherment system according to claim 1, wherein said predetermined tag is a frame tag.

[Claim 3]The web page decipherment system according to claim 1 or 2, wherein said URL detection means detects URL in syntax of <FRAMESRC="URL"> as URL in said frame tag.

[Claim 4]The web page decipherment system according to claim 1, wherein said sampling-of-text means extracts portions other than a portion surrounded by <> in source data memorized by said memory measure as a text part.

[Claim 5]The web page decipherment system according to claim 1 having further a voice output means which creates and outputs an audio signal corresponding to a text part extracted by said sampling-of-text means.

[Claim 6]The web page decipherment system according to claim 1, wherein an output sound

signal of said voice output means is supplied to telephone via a dial-up line.

[Translation done.]

* NOTICES *

JP0 and INPIT are not responsible for any damages caused by the use of this translation.

- 1.This document has been translated by computer. So the translation may not reflect the original precisely.
- 2.**** shows the word which can not be translated.
- 3.In the drawings, any words are not translated.

DETAILED DESCRIPTION

[Detailed Description of the Invention]

[0001]

[The technical field to which an invention belongs] This invention relates to the Web page decipherment system which decodes the text part of a Web (web) page.

[0002]

[Description of the Prior Art]WWW (World Wide Web) which is one of the information services of the Internet, Using the HTML file described in a language called HTML (Hyper Text Markup Language), and URL (Uniform Resource Locator) which is the identifiers of the preserving position of the file, via the Internet A character, an image, Multimedia information, such as a sound, can be referred to. It is a Web page which is formed on a display screen by processing an HTML file by the inspection software called a WWW browser. The WWW server associates and saves the HTML file by URL, and the side which provides the information on WWW operates according to server application. The source data which contain an HTML file by the side provided with information, i.e., a client, (client computer) from desired URL using a WWW browser. The Web page which accessed (for example, a graphics file and a voice file) via the Internet, and was formed by the file in sauce can be referred to on a display screen.

[0003]The document of HTML is usually formed from the text part which consists of text documents, and the display information bureau formed with a tag. A tag is a sign which makes <> a couple and the syntax of HTML is formed like <HTML>- </HTML> using a tag. Various display information, including the size of the character of the text part displayed on a Web page, the kind of font, its character color, the background color of a Web page, an image file name, an image position, etc., is shown in the portion surrounded by tag <>.

[0004]Thus, since the portion surrounded by tag <> in the document of HTML is the control information for displaying not the portion displayed but a document, if the portion surrounded by tag <> is removed, it is usually common to become a mere text document. The system

which, on the other hand, reads out the document of the Web page displayed on a WWW browser may be formed on the Internet. This is because the document of a Web page is read out and it sends out to telephone as an audio signal, when the terminal of a client is not a computer but the telephone connected to the public line, for example. Extracting the text data portion excluding the portion surrounded by tag <> out of the HTML file for read-aloud as Web page read-aloud, compounding the voice data corresponding to the character code of the text data portion, and outputting as a series of audio signals is performed.

[0005]

[Problem(s) to be Solved by the Invention]There is <FRAMESET>- </FRAMESET> in HTML as syntax for forming a frame on a Web page, for example, and the divided screen is obtained in the Web page using this. From the HTML file the syntax was described to be, another HTML file is further called for every split screen, and a document is displayed. That is, URL (an HTML file name is included) is specified for every split screen with the tag like <FRAME solvent refined coal="URL">, and the contents of the HTML file which exists in the specified field of URL are displayed.

[0006]However, there was a problem that a text part could not be correctly extracted to the document of HTML containing the tag which has description of URL like such a frame tag before. Then, the purpose of this invention is to provide the Web page decipherment system which can extract a text part correctly to the document of HTML containing the tag which has description of URL.

[0007]

[Means for Solving the Problem]A Web page decipherment system of this invention is provided with the following.

A means which is a Web page decipherment system which decodes a text part of an HTML document which constitutes a Web page, takes out source data of foundations which constitute a Web page from a storage area specified by predetermined URL (uniform resource locator), and is written in a memory measure.

A tag detection means which detects existence of a predetermined tag including a description part of URL out of basic source data.

A URL detection means to detect URL in the predetermined tag when existence of a predetermined tag is detected, A means which takes out source data from a storage area specified by URL detected by a URL detection means, and is written in a memory measure, and a sampling-of-text means to extract a text part from all the source data memorized by memory measure.

In the case of an HTML file in which basic source data contain a frame tag, from this composition, a text part of an HTML file stored in a field of URL described in that frame tag can also be extracted.

[0008]

[Embodiment of the Invention] Hereafter, the example of this invention is described in detail, referring to drawings. Drawing 1 shows the composition of the Web page decipherment system by that of this invention. In this system, WWW server 1 is a server which provides WWW as an information service, the HTML file is associated and saved by URL, and the graphics file and the voice file are also saved. WWW server 1 is connected to the Internet line network 2.

[0009] The CTI (Computer Telephony Integration) server 3 is connected to the Internet line network 2. CTI server 3 is connected also to the dial-up line network 4. Although two or more telephones are actually connected to the dial-up line network 4, the one telephone 5 is shown here. Telephones may be any of an ordinary phone machine, a public telephone, and a portable telephone. Although the office for dialups, such as a relay station and a base station, exists in the dial-up line network 4, it is not shown in a figure.

[0010] CTI server 3 is a server which provides with read-aloud of a Web page the user of the telephone containing the telephone 5. CTI server 3 is equipped with the Web page acquisition part 31, the sampling-of-text part 32, and the text read-aloud part 33. The Web page acquisition part 31 accesses WWW server 1, and acquires the source data of a Web page. The sampling-of-text part 32 analyzes the source data acquired by the Web page acquisition part 31, and extracts a text part. The text read-aloud part 33 creates a synthesized speech signal according to the character code of a text, and outputs the synthesized speech signal to telephone using the dial-up line network 4. The Web page acquisition part 31, the sampling-of-text part 32, and the text read-aloud part 33 are formed by the operation like the after-mentioned of the processor (not shown) of CTI server 3.

[0011] CTI server 3 has the memory storage 35, such as a hard disk, inside, and various data, such as source data, is memorized by the memory storage 35 so that it may mention later. In communication using the Internet line network 2 of WWW server 1 and CTI server 3 each, TCP/IP is used as a communications protocol, and the IP address is respectively assigned to WWW server 1 and CTI server 3. HTTP is used as a protocol of WWW. Although not illustrated, WWW server 1 and CTI server 3 are connected to the Internet line network 2 via the router.

[0012] Next, operation of this Web page decipherment system is explained. If a user telephones CTI server 3 from the telephone 5 and the talk state between the telephone 5 and CTI server 3 is established, as shown in drawing 2, CTI server 3, First, in order to acquire the source data of the Web page of the field specified by URL defined beforehand, it accesses to the URL (Step S1). Supposing this URL is in WWW server 1, WWW server 1 will read the source data which consist of files, such as an HTML file etc. of the field specified by URL, and will transmit to CTI server 3. The source data are the source data used as the foundations which constitute a Web page. The transmitted source data are supplied to CTI server 3 via the

Internet line network 2.

[0013]CTI server 3 stores in the memory storage 35 the source data sent from WWW server 1 (Step S2), and extracts a FRAME (frame) tag out of the stored source data (Step S3).

Although the extraction operation of this FRAME tag is mentioned later, URL contained in a FRAME tag is written in the memory storage 35 as a frame list.

[0014]CTI server 3 distinguishes whether extraction of the FRAME tag was performed as a result of execution of Step S3 (step S4). When extraction of a FRAME tag is not actually performed in step S4, it progresses to the below-mentioned step S7. On the other hand, when extraction of a FRAME tag is actually performed, In order to acquire the source data of the Web page of the field specified by URL of a frame list, it accesses to URL of the frame list (Step S5), and the source data sent from WWW server 1 are stored in the memory storage 35 (Step S6). Operation of WWW server 1 to access of Step S5 is the same as that of the case of access of Step S1. After execution of Step S6 progresses to Step S7.

[0015]CTI server 3 extracts text parts other than the portion surrounded by tab <> from the source data stored in the memory storage 35 in Step S7, reads out the extraction text data, and writes it in the memory storage 35 as data (Step S8). Then, a synthesized speech signal is created based on read-aloud data (step S9), and the synthesized speech signal is outputted to the telephone 5 (Step S10). since the read-aloud data written in the memory storage 35 is text data which consists of two or more character codes, the voice data corresponding to the character code group of the character codes of each and word unit is searched and obtained from the memory storage 35, and the synthesized speech signal which compounds these voice data and continues is created. A synthesized speech signal is supplied to the telephone 5 via the dial-up line network 4, it reads out from the receiver of the telephone 5, and a sound is outputted. The data table showing the relation between a character code and voice data in the memory storage 35 is memorized beforehand.

[0016]Next, it explains, referring to the flow chart of drawing 3 for the FRAME tag extraction operation out of the source data in the above-mentioned step S3. CTI server 3 searches character string <FRAME solvent refined coal out of the source data memorized by the memory storage 35 (Step S11), and distinguishes whether character string <FRAME solvent refined coal exists in source data (Step S12). That is, an HTML file is contained in the source data written in the memory storage 35, and it is distinguished whether frame settings are performed by the document which the HTML file shows. Since the syntax of <FRAME solvent refined coal="URL"> exists if character string <FRAME solvent refined coal exists, A read position is moved to the position of the following character = (Step S13), the character string surrounded by subsequent "", i.e., URL, is further read in source data, and the URL is written in the frame list formed in the memory storage 35 (Step S14). Therefore, URL which shows the existence position of the HTML document contained in a frame is written in a frame list. It

distinguishes whether search of character string <FRAME solvent refined coal was completed from all the files of source data after execution of Step S14 (Step S15), when the search is not completed, it returns to Step S11, and the above-mentioned step operation is repeated.

[0017]Subsequently, it explains, referring to the flow chart of drawing 4 for the extraction operation of the text data in the above-mentioned step S7. CTI server 3 acquires the character code for one character sequentially from the head of one file in the source data stored in the memory storage 35 (Step S21) -- the character code -- a character -- < -- it is distinguished whether it is shown or not (Step S22). the acquired character code -- a character -- < -- when shown, tag flag F_{TAG} is made equal to one (Step S23). the acquired character code -- a character -- < -- when not shown, it is distinguished whether tag flag F_{TAG} is [one] equal (Step S24). It is a flag which tag flag F_{TAG} is set as 1 in the portion surrounded by <> in the HTML file, and is set as 0 in the other portion, and the initial value is 0. Distinction of Step S24 is performed even after execution of Step S23.

[0018]As a result of distinction of Step S24, when tag flag F_{TAG} is made equal to one, the character code acquired at Step S21 distinguishes whether character > is shown (Step S25). Since it is the end of a tag when the acquired character code shows character >, tag flag F_{TAG} is made equal to zero (Step S26). On the other hand, as a result of distinction of Step S24, since it is text parts other than the tag portion surrounded by <> when tag flag F_{TAG} is made equal to zero, the acquired character code is read out, and it stores in the memory storage 35 so that it may be made to contain in data (Step S27).

[0019]After Step S26 or execution of S27 distinguishes whether acquisition of the character code for one character was completed from all the source data (Step S28), when the acquisition is not completed, it returns to Step S21, and it repeats the above-mentioned step operation. Therefore, since the text part of the HTML file stored in the field of URL which was described in the FRAME tag in the case of the HTML file in which basic source data contain a FRAME tag can also be extracted according to this system, It can read out without leaving the text part of the Web page displayed on a WWW browser.

[0020]In the above-mentioned example, although the case where a frame tag was used was explained, this invention is applicable also to the tag which has description of URL other than a frame tag. This invention can be applied also when extracting and reading out a text part from the HTML file containing script languages, such as JavaScript. In HTML, in order to extend a function, the tag which enables it to perform script languages, such as JavaScript, on a Web page is also prepared. For example, it is formed by syntax like <SCRIPT LANGUAGE="JavaScript">- </SCRIPT>. Therefore, as described above, the portion of the range of <SCRIPT LANGUAGE="JavaScript">- </SCRIPT> is disregarded and the other text

part is extracted.

[0021]The system of this invention is applicable also to the file using the language to which the file which used HTML was made to extend not only a case but HTML. Although all the pages that can be perused in a WWW browser were called the homepage in Japan, since the homepage was originally a basic page of two or more Web pages which constitute one information group, it was indicated with the Web page that misunderstanding was not invited here.

[0022]

[Effect of the Invention]Like the above, a text part can be correctly extracted in the Web page decipherment system of this invention to the document of HTML containing the tag which has description of URL. Therefore, it can read out, without leaving the text part of a Web page, when reading out a Web page.

[Translation done.]

* NOTICES *

JP0 and INPIT are not responsible for any damages caused by the use of this translation.

- 1.This document has been translated by computer. So the translation may not reflect the original precisely.
- 2.**** shows the word which can not be translated.
- 3.In the drawings, any words are not translated.

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]

[Drawing 1]It is a block diagram showing the composition of the Web page decipherment system by this invention.

[Drawing 2]It is a flow chart which shows operation of the CTI server in the system of drawing 1.

[Drawing 3]It is a flow chart which shows the FRAME tag extraction operation out of source data.

[Drawing 4]It is a flow chart which shows the extraction operation of text data.

[Description of Notations]

- 1 WWW server
- 2 Internet line network
- 3 CTI server
- 4 Dial-up line network
- 5 Telephone

[Translation done.]

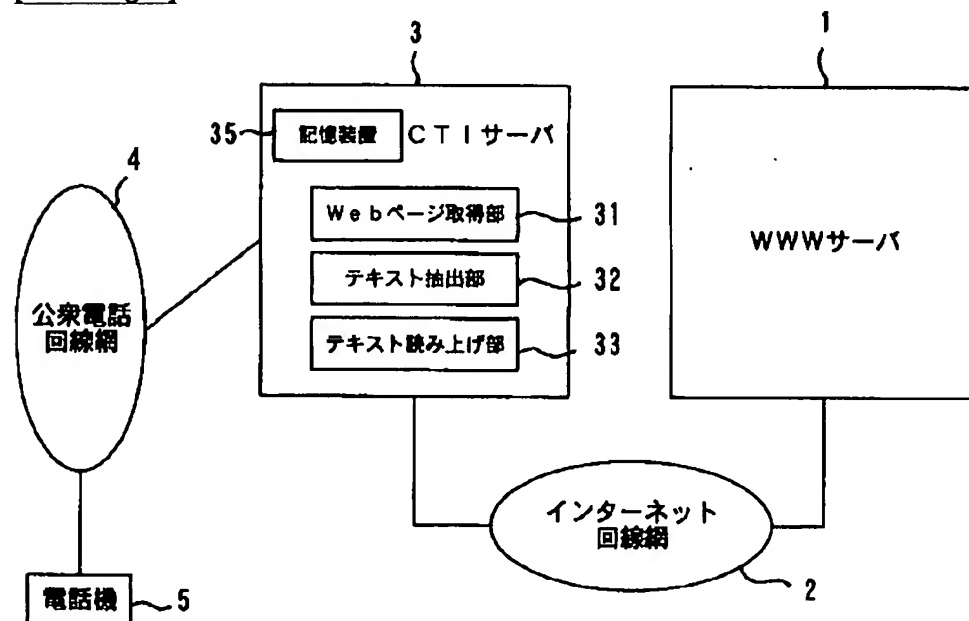
* NOTICES *

JPO and INPIT are not responsible for any damages caused by the use of this translation.

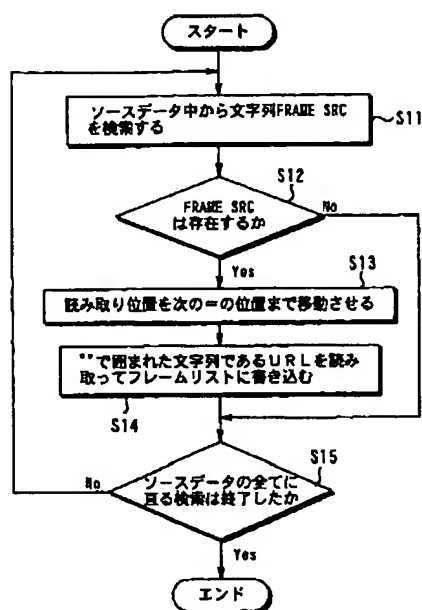
- 1.This document has been translated by computer. So the translation may not reflect the original precisely.
- 2.**** shows the word which can not be translated.
- 3.In the drawings, any words are not translated.

DRAWINGS

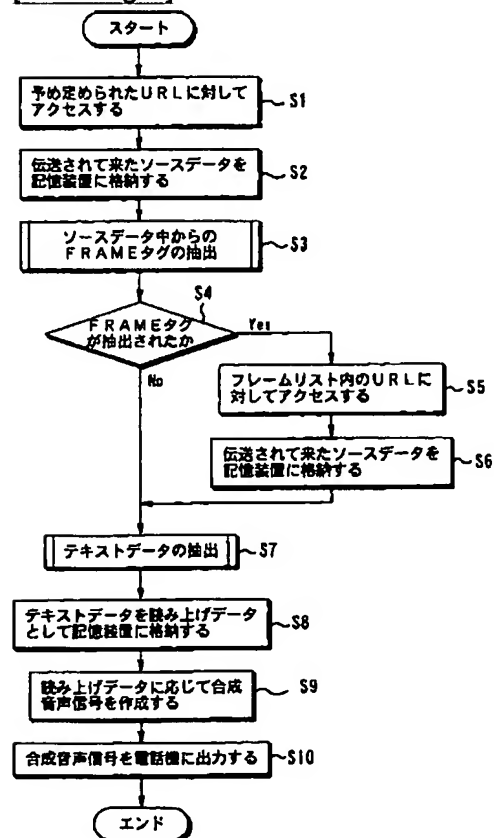
[Drawing 1]



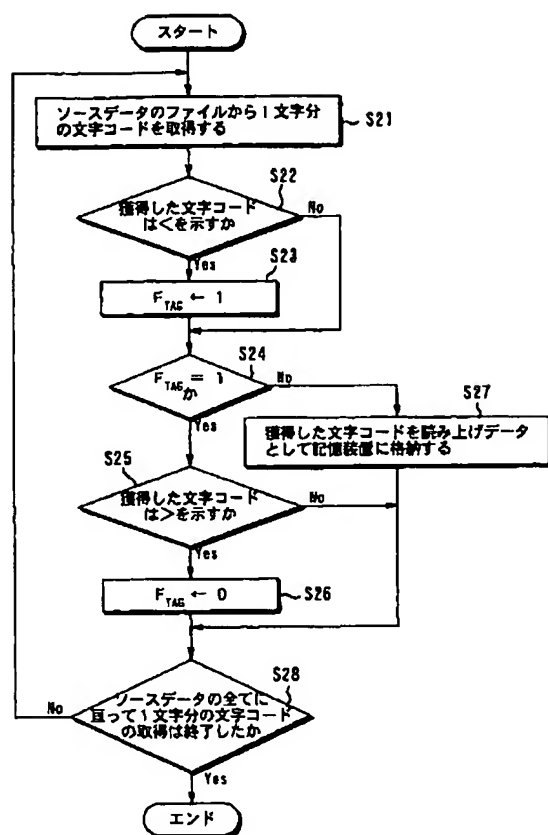
[Drawing 3]



[Drawing 2]



[Drawing 4]



[Translation done.]